面向移动 App 流量的多特征集合集成聚类方法研究与应用 *

吴志敏 [1]，刘　珍 [1†]，王若愚 [2,3]，陈洁桐 [1]

(1. 广东药科大学 医药信息工程学院，广州 510006; 2. 华南理工大学 信息网络工程研究中心，广州 510041; 3. 广东省计算机网络重点实验室，广州 510041)

摘　要：针对移动互联网流量识别问题，基于多项性能评估指标，分析 K-均值和谱聚类算法在不同特征集合或不同识别目标的流量数据集上的聚类性能；并提出基于多特征集合的集成聚类方法。比较分析实验表明，相同聚类方法在不同特征集合或不同识别目标数据集上性能有所不同，集成聚类方法能够有效提高利用单个特征集合的聚类方法的性能。进一步将集成聚类方法应用于 App 关联分析，分析结果可为移动 App 的划分和用户行为分析提供客观依据。

关键词：移动 App 流量；流量统计特征；集成聚类；流量识别

中图分类号：TP393　　doi: 10.3969/j.issn.1001-3695.2018.04.0250

Research and Application of multi-feature sets based ensemble clustering method for mobile App traffic

Wu Zhimin[1], Liu Zhen[1†], Wang Ruoyu[2, 3], Chen Jietong[1]

(1. *School of Medical Information Engineering Guangdong Pharmaceutical University, Guangzhou* 510006, *China*; 2. *Information & Network Engineering & Research Center, South China University of Technology, Guangzhou* 510041, *China*; 3. *Communication & Computer Network Lab of Guangzhou* 510041, *China*)

Abstract: To handle the mobile traffic identification problem, based on multiple performance evaluation metrics, this paper analyzed the performance of K-Means and Spectral Clustering algorithms on the data sets characterized by different feature sets or labeled with different class set, and proposed an ensemble clustering method from the aspects of combining the clustering results on the data sets with different feature sets. Experimental results show that the performance of the same clustering algorithm is different on the data sets with different feature sets or traffic classes, and the ensemble clustering method is able to improve the overall clustering performance. Further, this paper applies the ensemble clustering method on the correlation analysis of mobile apps, and the results can support the decision on grouping apps and analyzing user behaviors.

Key words: mobile app traffic; traffic statistics features; ensemble clustering; traffic identification

0　引言

近年来，随着移动互联网与智能终端设备的快速发展，用户可以随时随地访问互联网。成千上万的智能手机应用每天产生海量数据，移动互联网流量数据日益庞大。网络流量是记录和反映网络及其用户活动的重要载体。通过网络流量识别，可以间接地掌握互联网的使用情况，从而为网络运营、监控和测量方面提供辅助决策[1, 2]。

基于机器学习的流量识别方法成为近年来的研究热点。早期，在传统互联网流量数据上，徐鹏等人[2]的实验比较分析表明支持向量机比朴素贝叶斯的流量分类性能更加稳定；Soysal 等人**错误!未找到引用源。**的比较实验结果表明决策树比贝叶斯网络和多层感知器在流量分类方面具有更高的准确性和有效性。后续多种互联网流量识别方法被提出，主要关注于在线流量分类问题**错误!未找到引用源。**、不平衡分类问题**错误!未找到引用源。**、分类鲁棒性[4,6,7]等。聚类方法的优点是无需有标记数据参与模型训练[8,9]。在互联网流量聚类方法研究方面，多种聚类方法被用于互联网流量识别，例如 K 均值、高斯混合模型和谱聚类[10,11]，DBSCAN **错误!未找到引用源。**等。近期，鲁刚等人[12]利用前 *N*（*N*=1，…，10）个报文大小的特征建立基聚簇模型，然后

收稿日期：2018-04-04；修回日期：2018-05-29　**基金项目：**国家自然科学基金资助项目（61501128）；广东省自然科学基金资助项目（2017A030313345）；国家级大学生创新创业训练计划项目（201710573005）；中央高校基本业务费资助项目（x2rj/D2174870）

作者简介：吴志敏（1995-），男，广东梅县人，学士，主要研究方向为数据挖掘与机器学习；刘珍（1986-），女（通信作者），四川宜宾人，博士，主要研究方向为互联网流量分类、机器学习（liu.zhen@gdpu.edu.cn）；王若愚（1977-），男，江西乐平人，博士，主要研究方向为计算机网络、模式分类；陈洁桐（1995-），女，广东汕尾人，学士，主要研究方向为软件工程．

利用基聚簇模型进行聚类，基于聚类概率作为新的特征建立新数据集，并利用有监督学习方法 SVM 做最终决策，但是实验数据仍然是传统互联网流量。在移动互联网流量上，已有文献主要关注基于载荷的 App 识别错误!未找到引用源。和基于机器学习的 App 行为识别**错误!未找到引用源。**，文献**错误!未找到引用源。**基于聚类方式识别移动互联网流量的 P2P、WEB 等服务类型，基于聚类方法的移动 App 流量识别研究较为缺乏错误!未找到引用源。。

已有互联网流量识别相关研究工作面临如下问题：a）各文献采用了不同的特征集合[18]，如单向流特征集合[2,19]、双向流特征集合[20,21]；各实验数据集的识别目标也有所不同，例如 App 级别（微信、QQ 等）错误!未找到引用源。、用户行为级别（文本聊天、视频通话等）错误!未找到引用源。，各文献的实验结果不能直接进行比较;b）为管理大量的 App，通常根据主观意识进行 App 类别划分，建立粗粒度识别目标，但是这种方式存在主观随意性，缺乏客观依据。

针对上述问题，本文主要贡献如下：

a）基于 Mobilegt 系统**错误!未找到引用源。**，采集移动互联网流量数据集，在数据集上提取四种不同的流量特征集合，开展 App 级别和上网行为（Behavior）级别的流量类别标记工作，为本文的聚类方法研究提供数据基础。

b）利用多项性能评估指标，在不同特征集合、不同粒度类别标签的流量数据集上，比较分析各聚类方法、特征集合的性能等。

c）为综合利用不同角度特征集合的优势，提出集成聚类方法，进一步提高单特征集合建立的聚类模型的流量识别性能；

d）基于集成聚类方法，提出移动 App 相似度评价指标，此相似度分析结果为 App 归类（如社交类、视频类等）提供客观建议，并辅助用户上网行为分析。

1　集成聚类算法

将机器学习算法用于网络流量识别，需要首先对原始报文建立网络流，并提取流统计特征（如报文大小、流持续时间等统计特征），然后建立特征向量描述的流样本集合，将其作为机器学习算法的输入，训练识别模型。已有研究表明，不同角度的流统计特征集合**错误!未找到引用源。**已被提出，并用于网络流量识别。聚类算法比较实验结果（详情见 3.2 小节）显示不同角度的特征集合可能有各自的优势，不分伯仲。受此启发，结合集成学习模型的特点，集成各特征集合描述的流量数据集上的聚类结果，可进一步提高聚类方法的性能。据此，本文提出基于多个特征集合的集成聚类算法（**m**ulti-**f**eature sets based **e**nsemble **c**lustering, MFEC）。

1.1　基本概念

为方便理解本文的网络流量识别工作，本小节首先给出相关的基本概念。

a)网络流量识别：将网络 IP 报文映射为流量类别（例如移动 App、用户行为等）。

b)网络流：在一定时间间隔内（如 300s），具有相同五元组{源 IP、源端口、目的 IP、目的端口、传输层协议}的 IP 报文组成。

c)流统计特征：在组成网络流的 IP 报文上，提取报文大小、报文到达时间间隔等统计值。

1.2　基于多特征集合的集成聚类算法（MFEC）

假设有 m 个特征集合，对网络流量数据 S 特征化后，建立 m 个样本集合{S_1, .., S_m}，S_i={($\mathbf{x}^i_1$, y_1), ($\mathbf{x}^i_2$, y_2), ..., ($\mathbf{x}^i_n$,y_n)}，$\mathbf{x}^i_j$ 表示利用第 i 个特征向量描述第 j 条网络流建立的流样本。集成聚类方法的伪代码如算法 1。主要步骤包括：

a）在 m 个样本集合上，利用基础聚类算法（例如 K 均值），分别训练聚类模型{f_1, f_2, …, f_m}。

b）为了处理集成聚类的不一致问题，利用每个聚类模型 f_i，建立一个流样本之间的关联矩阵 $\mathbf{M}$，$M[i][j][k]$记录了 f_i 是否将 $\mathbf{x}^i_j$ 和 $\mathbf{x}^i_k$ 划分到一个聚簇中（若相同，取值为 1，否则为 0）;

c）各关联矩阵相乘，最后将两两关联的样本划分到同一个聚簇。

这意味着，仅当 m 个聚类模型将某些样本划分到同一个聚簇时，它们最终才在同一个聚簇。例如有 5 个流样本，表示为 $\mathbf{x}_1$, $\mathbf{x}_2$, $\mathbf{x}_3$, $\mathbf{x}_4$, $\mathbf{x}_5$，有 2 个聚类模型，分别开展聚簇后获得的样本关联矩阵为 $\mathbf{M}_1$ 和 $\mathbf{M}_2$. 基于 $\mathbf{M}_1$ 可得{x_1, x_2, x_3}; {x_4, x_5}两个簇；基于 $\mathbf{M}_2$ 可得{x_2, x_3}; {x_1, x_4, x_5}两个簇。集成两个聚簇结果后的聚簇由 $\mathbf{M}$=$\mathbf{M}_1$·$\mathbf{M}_2$ 获得。集成后得到{x_1}; {x_2, x_3}; {x_4, x_5}三个簇。

$$\mathbf{M}_1=\begin{bmatrix}1&1&1&0&0\\1&1&1&0&0\\1&1&1&0&0\\0&0&0&1&1\\0&0&0&1&1\end{bmatrix}\quad \mathbf{M}_2=\begin{bmatrix}1&0&0&1&1\\0&1&1&0&0\\0&1&1&0&0\\1&0&0&1&1\\1&0&0&1&1\end{bmatrix}\quad \mathbf{M}=\mathbf{M}_1*\mathbf{M}_2=\begin{bmatrix}1&0&0&0&0\\0&1&1&0&0\\0&1&1&0&0\\0&0&0&1&1\\0&0&0&1&1\end{bmatrix}$$

算法 1　MFEC 集成聚类算法

输入　S_1, …, S_m

输出　样本聚簇标签

```
for i = 1 to m
f_i = BasicClustering(S_i)//在每个数据集上训练一个聚类模型
end
for i = 1 to m
    for j = 1 to n
      for k = 1 to n
    if f_i(x_j) == f_i(x_k)
      M[i][j][k] = 1; //构建关联矩阵 M
    end if
      end for
    end for
end for
for i = 2 to n
  M[1] = M[1]·M[i];
end for
```

labels[] = *getCluster*(*M*[1][][]);//基于关联矩阵 **M** 获得每个样本簇号

2　移动互联网流量数据采集与预处理

2.1　流量数据采集

基于 Mobilegt 系统**错误!未找到引用源。**的实验数据采集环境部署如图 1 所示。Mobilegt 包括客户端 MgtClient 应用，服务器端 MgtServer 程序。在移动终端设备上安装 MgtClient，将 MgtServer 程序部署到服务器端。用户在移动终端开启 MgtClient，并点击连接按钮，启动 VPN 服务和 Socket 数据采集程序；然后用户像往常一样使用其它应用，例如微信、微博、浏览网页等。移动端产生的流量会重路由到服务器端，在服务器端 MgtServer 程序采集客户端发出和接收的所有网络流量数据。MgtClient 采集所有网络会话信息，即：五元组（源 IP、源端口、目的 IP、目的端口和传输层协议）与 App 的映射关系，并记录到 Socket 文件。当移动端结束数据采集，MgtClient 将 Socket 数据发送到服务器端，MgtServer 程序接收 Socket 文件，并基于 Socket 文件对采集的网络流量数据进行 App 标记工作。Socket 文件记录了客户端网络流量的真实 App 信息。因此，MgtServer 可以利用 Socket 文件对网络流量进行 100%准确率的类别标记工作。

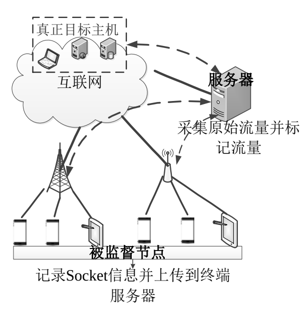


图 1　数据采集环境

2.2　流量数据预处理

在采集的移动互联网流量报文数据上，首先组流并提取流统计特征，建立流样本集合。已有多种流统计特征被用于流量识别，最常用的是基于报文大小和报文到达时间间隔的统计计算。统计的网络流对象又分为双向流和单向流两种。相关定义如下：

假设某移动设备（IP_1，$Port_1$）与某服务器（IP_2，$Port_2$）利用传输层协议 *pro* 进行通信。在一定时间间隔内，它们之间通信的 IP 报文组成了网络流。

定义 1　单向流。由单个方向的 IP 报文组成，OUT 方向的网络流表示为{*Pkt*| srcIP(*Pkt*)=IP_1 & dstIP(*Pkt*)= IP_2 & srcPort(*Pkt*) = $Port_1$& dstPort(*Pkt*) = $Port_2$ & Protocol(*Pkt*) = *pro*}；IN 方向的网络流表示为{*Pkt*| dstIP(*Pkt*)=IP_1 & srcIP(*Pkt*)= IP_2 & dstPort(*Pkt*) = $Port_1$ & srcPort(*Pkt*) = $Port_2$ & Protocol(*Pkt*) = *pro*}。

定义 2　双向流。由两个方向的 IP 报文组成，表示为{*Pkt*| (srcIP(*Pkt*)=IP_1 & dstIP(*Pkt*)= IP_2 & srcPort(*Pkt*) = $Port_1$ & dstPort(*Pkt*) = $Port_2$) ‖ (dstIP(*Pkt*)=IP_1 & srcIP(*Pkt*)= IP_2 & dstPort(*Pkt*) = $Port_1$ & srcPort(*Pkt*) = $Port_2$) & Protocol(*Pkt*) = *pro*}。

基于以上定义，本文提取的四种流统计特征描述如下：

1）单向流统计特征（UniDirection）**错误!未找到引用源。**

IN 总报文数，IN 字节数，IN 报文大小(最小、最大、平均、标准差、峰度、偏度、标准误差)，IN 报文到达时间间隔(最小、最大、平均、标准差)，IN 流持续时间；OUT 总报文数，OUT 字节数，OUT 报文大小(最小、最大、平均、标准差、峰度、偏度、标准误差)，OUT 报文到达时间间隔(最小、最大、平均、标准差)，OUT 流持续时间。

2）双向流统计特征（BiDirection）

总报文数，字节数，报文大小(最小、最大、平均、标准差、峰度、偏度、标准误差)，报文到达时间间隔(最小、最大、平均、标准差)，流持续时间。

3）前 *k* 个报文大小分布（PS）[24]

前 *k* 各报文大小，分别表示为{$ps_1, ps_2, \ldots, ps_k$}。

4）前 *k* 个报文大小分布-映射（PS-mapped）[24]

将前 *k* 个报文的报文大小映射为 4 个数值｛1, 2, 3, 4｝，新的特征表示为{v_1, v_2, …, v_k}。IN 方向的报文大小的映射如式（1），OUT 方向的报文大小的映射如式（2）所示。

$$v_i=\begin{cases}1 & ps_i=[0,150]\\ 2 & ps_i=[150,700]\\ 3 & ps_i=[700,1300]\\ 4 & ps_i=[1300,1500]\end{cases} \tag{1}$$

$$v_i=\begin{cases}-1 & ps_i=[0,150]\\ -2 & ps_i=[150,700]\\ -3 & ps_i=[700,1300]\\ -4 & ps_i=[1300,1500]\end{cases} \tag{2}$$

2.3　实验数据

2.3.1 SAD（specific application data，特定应用数据）

为了验证不同特征集合在不同类别标签数据集上的性能，本文针对三种常用社交应用（QQ、微信、微博），采集了 SAD 数据集。采集过程为：用户根据规定的 App 和上网行为（例如微信视频通话、微信文本通话等），运行特定的 App 并执行相应的行为，每个应用持续时间大概 20 min。

表 1　App 类别标签数据集

App 类别	流数目	报文数目	字节数目/MB
QQ	206	1 633 031	1 127.587
WeChat	272	475 412	188.930
Weibo	336	122 287	96.889
总数目	814	2230730	1413.406

在运行过程中 Mobilegt 系统采集原始报文数据、组流、提取特征、并根据 Socket 信息标记网络流样本的 App 类别标签。根据用户在规定时间内运行特定 App 的上网行为记录，以人工标记方式标记流样本的 Behavior 类别标签。这意味着，上述采集的原始流量数据上赋予了两种类别标签（App 类别标签和 Behavior 类别标签），两种类别标签的流量数据在类间的分布如表 1 和表 2 所示。两个数据集的数据来源相同，只是流样本的

标签类型不同，两个数据集的总体流数目、报文数目和字节数目相同。

表 2　Behavior 类别标签数据集

Behavior 类别	流数目	报文数目	字节数目/MB
audiochat	13	248 428	37.607
browse	361	107 828	85.090
chat	267	243 984	203.960
post	160	1 048 508	850.453
videochat	13	581 982	236.296
总数目	814	2230730	1413.406

2.3.2 MAD（more applications data，多应用数据）

SAD 数据集涉及到人工标记，采集的数据有限。为了验证集成聚类方法在更多 App 流量数据集上的性能。本文另外采集了 MAD 数据集，此数据集来自于多个用户，用户根据自己的意愿开启 mobilegt 终端应用，mobilegt 服务器端自动采集和标记流量数据，采集时间为 2017 年 1 月，然后抽取流行的 9 个 App 进行实验。MAD 数据的类间分布情况如表 3 所示，相比 SAD，其具有更多的不同 App 标签。在这组数据上，主要用于验证集成聚类方法的性能，并分析集成聚类应用情况。

表 3　MAD 数据集

App 类别	流数目	报文数目	字节数目/MB
Browser	4 000	469 578	281.202
JdShop	1 209	216 806	147.358
MgTV	4 000	3 153 559	2 208.004
QQ	4 206	3 874 852	2 774.752
VipShop	1 745	433 065	253.586
WeChat	3 684	1 037 697	651.094
Weibo	4 336	1 758 350	1 482.330
YahooMail	2 443	111 108	51.998
Youku	3 054	14 63 226	1 078.451

3　聚类方法比较实验

3.1　实验设计

本文实验首先比较不同的聚类方法在不同特征集合和不同类别标签流量数据集上的聚类性能，然后再实验比较本文提出的 MFEC 集成聚类算法的性能。目前尚未查阅到相关文献比较分析不同流统计特征集合的性能。在 SAD 数据集上，提取了 2.2 小节的 4 个特征集合，并且 SAD 标记了 2 种类别标签，从而可得到 8 个不同的数据集。在各数据集上，利用 K 均值和谱聚类方法开展聚类，然后讨论和分析在不同 k 值条件下，多项聚类性能评估指标的变化情况。考虑到不同特征的量纲对聚类结果的影响，实验采用 Min-Max 方式对所有特征进行归一化处理。每个聚类方法在各数据集上独立重复执行 20 次，实验结果为 20 次的平均。为了从不同角度分析聚类方法的性能，本文采用了以下三种评估指标。

1）信息熵

本文利用信息熵评估每个聚簇的纯度。给定某个随机变量 X，信息熵定义为式（3），信息熵评价随机变量取值的离散程度。为评估聚簇中 App 分布的离散程度，式（3）中的 p_i 表示第 i 个 App 在聚簇中的百分比。信息熵取值越小，表示聚簇的纯度越高。

$$Entropy(X) = -\sum_{i=1}^{n} p_i \log_2 p_i \qquad (3)$$

2）轮廓系数

轮廓系数结合了簇内紧密度和簇间分离度两种因素，如式（4）所示。其中，a 是该样本与同簇其它样本的平均距离，b 是与其距离最近的它簇样本的平均距离。$s\in[-1, 1]$，s 越接近于 1，聚类效果越好。

$$s = \frac{b-a}{\max(a,b)} \qquad (4)$$

3）CH 分数

CH 分数与轮廓系数的区别在于，CH 分数是通过计算聚簇内各点与其中心的距离平方和表示簇内紧密度，计算各聚簇中心点与数据集中心点的距离平方和表示簇间分离度，如式（5）所示。其中，tr 表示矩阵的迹，U_k 为簇间分离度矩阵，W_k 为簇内紧密度矩阵。n 为样本数，k 为聚簇数。CH 分数越大代表聚簇自身越紧密，聚簇之间越分散。

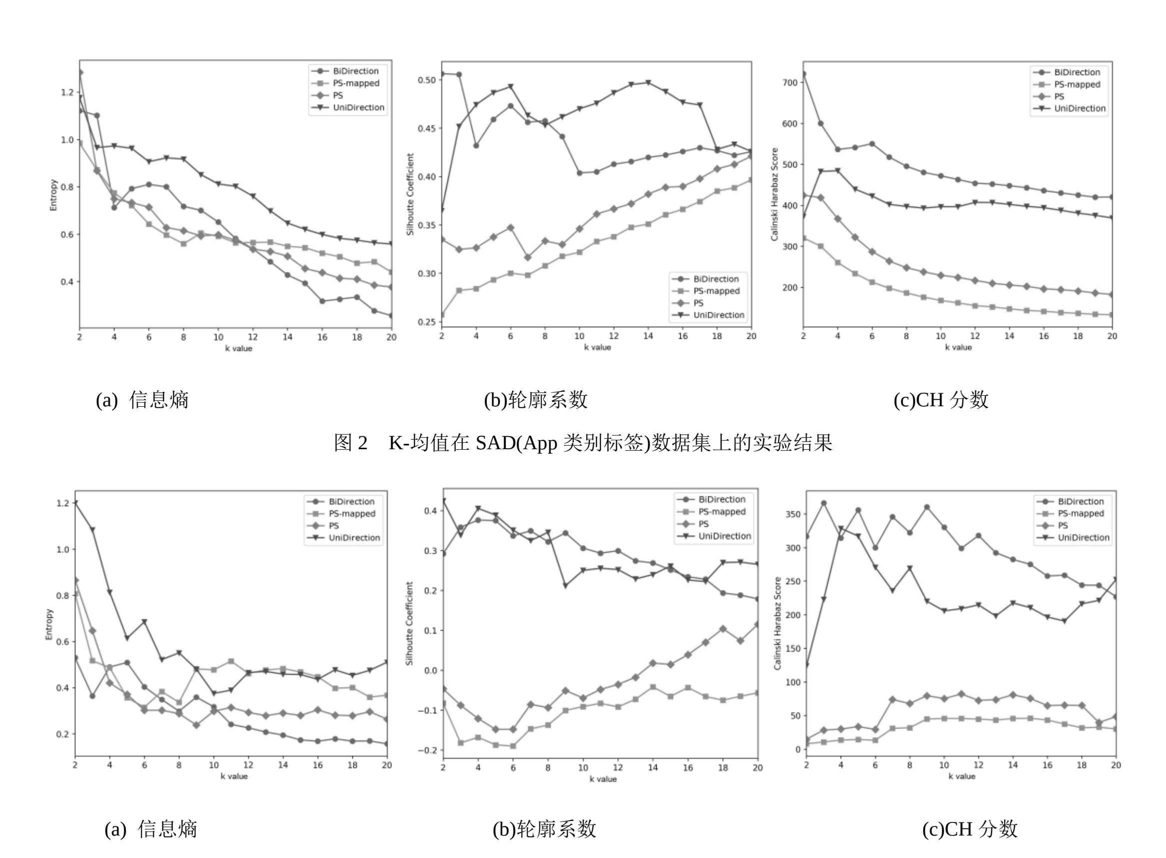
$$c = \frac{tr(U_k)\cdot(n-k)}{tr(W_k)\cdot(k-1)} \qquad (5)$$

3.2　K-均值与谱聚类在不同数据集上的性能分析

本小节旨在分析 K-均值和谱聚类算法在不同特征集合和不同类别标签的流量数据集上的性能，探究 a）聚类算法在不同特征集合的流量数据集上的性能是否稳定；b）聚类算法在不同类别标签的流量数据集上的性能是否稳定；c）K-均值与谱聚类算法比较，哪种算法在流量数据上的性能更优；d）哪一类特征集合在流量数据上表现更优。

3.2.1 在 SAD(App 类别标签)数据集上的性能

K-均值和谱聚类算法在 App 类别标记数据集上的实验结果分别如图 2 和图 3 所示。a）信息熵,随着 K 值增加，K-均值和谱聚类的信息熵都随之降低，K-均值表现相对稳定，波动次数更少,并且双向流特征集的信息熵最小;b）轮廓系数,谱聚类的效果劣于 K-均值,对于表现较好的单向流与双向流特征集，谱聚类从 K 取初值到逐渐增大的过程中,其轮廓系数基本维持在 0.4 以下，而 K-均值却能够稳定在 0.4 以上;c）CH 分数：对于 4 个特征集合表示的数据集，K-均值的 CH 分数最小值分别约为 480、420、220、160，皆大于谱聚类的最大值分别约 370、320、80、45；与轮廓系数类似，双向流特征和单向流特征获得更好的性能。



(a) 信息熵　　(b)轮廓系数　　(c)CH 分数

图 2　K-均值在 SAD(App 类别标签)数据集上的实验结果

(a) 信息熵　　(b)轮廓系数　　(c)CH 分数

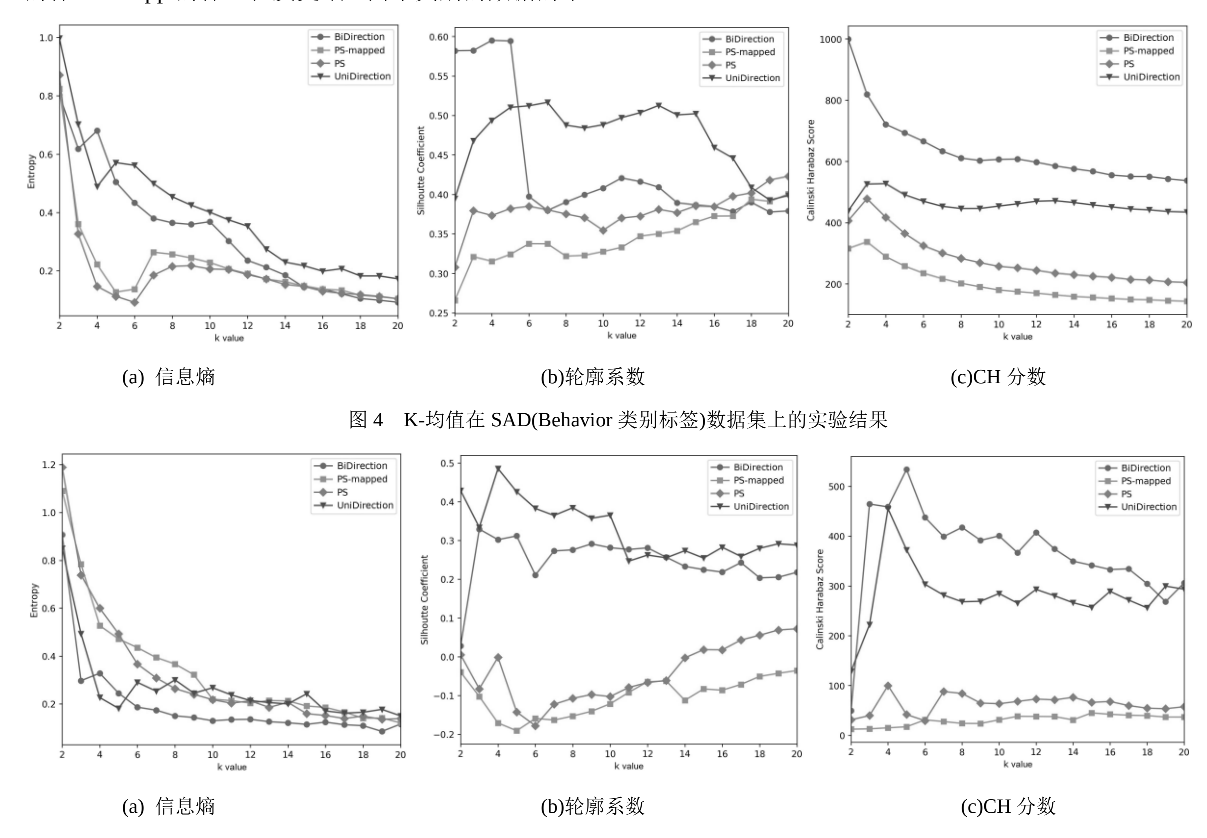
图 3　谱聚类在 SAD(App 类别标签)数据集上的实验结果

3.2.2 在 SAD(behavior 类别标签)数据集上的性能

K-均值和谱聚类在 Behavior 类别数据集上的实验结果分别如图 4、5 所示。

a）信息熵。与 App 标签数据集上的实验结果相比，两种聚类算法在 Behavior 标签数据集上的性能更优，这是由于 Behavior 的标签比 App 的标签粒度更细，同个类别的数据的聚集程度更高，更易聚类；当 K 值增加到最大 20 时 4 种特征集合的性能相差不多。

b）轮廓系数与 CH 分数：类似 App 数据的性能，K-均值比谱聚类性能更优；在特征集合方面，单向流特征和双向流特征集合性能更优。



(a) 信息熵　　(b)轮廓系数　　(c)CH 分数

图 4　K-均值在 SAD(Behavior 类别标签)数据集上的实验结果

(a) 信息熵　　(b)轮廓系数　　(c)CH 分数

图 5　谱聚类在 SAD(Behavior 类别标签)数据集上的实验结果

基于上述实验结果分析，得出以下结论：

a）在聚类方法比较方面，K-均值的聚类纯度（信息熵）接近谱聚类的性能情况下，其聚类结果的聚簇内更紧密，簇间距离更远。并且 K-均值的计算开销低于谱聚类，K-均值更适用于移动互联网流量识别。

b）在 4 个特征集合比较方面，单向流特征和双向流特征集总是表现更优。这是由于这两类特征的信息量更丰富，对报文大小和报文到达时间间隔进行了不同角度的统计计算。但是，两者之一没有总是表现最优，各有不同的适应场景。本文提出的 MFEC 方法，可集成两种特征集合表征的数据集上的聚类结果，提高单个特征的聚类性能。

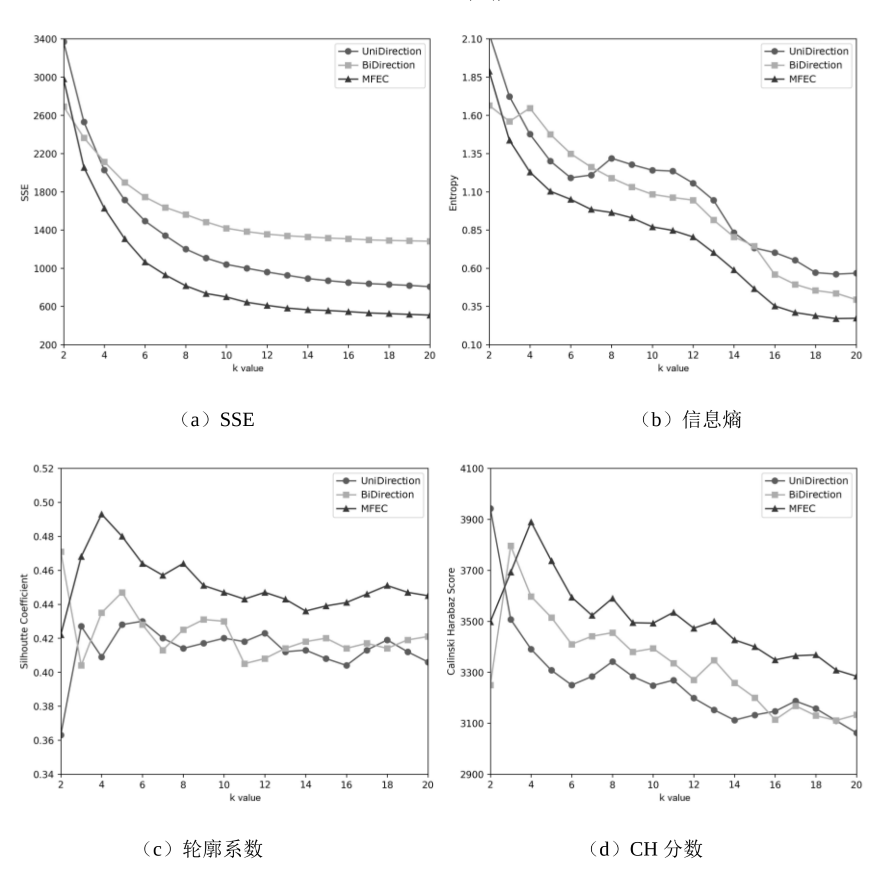
实验结果还表明，在 App 目标数据上比在 Behavior 目标数据上的聚类更难，这是由于 App 包括了多种行为，不同 App 之间还存在相似的行为。第 3.3 小节将在较难识别的多个 App 标记的数据集上验证 MFEC 的性能。

3.3　MAD 数据集上 MFEC 方法性能验证分析

根据第 3.2 小节实验结果，K-均值相对更适用于移动互联网流量识别，本小节将采用 K-均值作为基础算法，单向流特征（UniDirection）和双向流特征（BiDirection）作为特征集合，在规模更大、App 类别更多的 MAD 数据集上验证 MFEC 方法的性能。在性能评价指标方面，采用上述三项指标之外，SSE（sum of squared errors）和 App 识别准确率也将作为评价指标。为了缓解类间样本不平衡对聚类实验性能的影响，每个应用抽样了 4 000 条网络流，不足 4 000 的则全部抽样。

3.3.1 MFEC 实验性能分析

MAD 数据集上聚类评估结果如图 6 所示，其中，UniDirection 和 BiDirection 表示采用相应特征集合的 K-均值，MFEC 表示本文的集成聚类方法。实验结果表明，随着 *K* 值的增加，K-均值和集成聚类产生的 *SSE* 和信息熵呈现较为平稳的下降趋势；轮廓系数在整体上是上升的，而 CH 分数是下降的，且随着 *K* 值的增加而渐渐趋向平稳。其原因可能是随着 *K* 值增加，属于相同类别的流样本划分到多个聚簇，簇间的距离可能降低。总体上，集成聚类的性能总是保持最优，实验结果验证了 MFEC 方法可以进一步提高利用单个特征集合的 K-均值的性能。



（a）SSE　　（b）信息熵

（c）轮廓系数　　（d）CH 分数

图 6　MFEC 与 K-均值的四项指标性能比较

为验证各聚类方法在 App 识别准确率方面的性能，本文借鉴文献[9]的做法，对于每个聚簇可将样本数最多的 App 标签作为这个簇的所有样本的预测类别标签，从而计算识别准确率。识别准确率可定义为，对于给定的应用样本识别出的应用流数与总流数之比，定义如下：

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{6}$$

图 7 显示了 K-均值与 MFEC 对 App 的识别准确率。实验结果表明，当 $2 \le K < 7$ 时，K 均值在整体上比 MFEC 的识别准确率要高。随着 *K* 值增大，样本得以更细的划分，因而各自的准确率也逐渐增加。其中 MFEC 的提升尤为明显，并明显地超过了 K-均值。MFEC 的结果表明，平均性能可达到 70%以上的 App 识别准确率。

基于多特征集的 MFEC 的性能优于 K-均值聚类。其主要

原因可能是，集成聚类能够综合利用在不同特征集上建立的多个聚类模型所得结果，在一定程度上起到"集优"的效果。另外，采用集成聚类可以帮助提高 App 识别准确率。当 *K* 取值更大时，集成聚类更能显出它的这一优势。

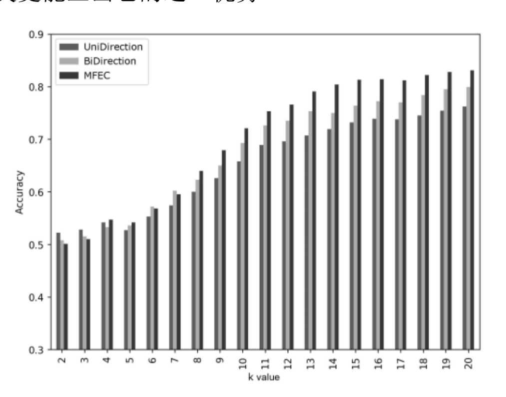


图 7　MFEC 与 K 均值的识别准确率比较结果

3.3.2 基于集成聚类的 App 相似度分析

目前，已有大量移动 App，而将机器学习算法用于大量 App 分类的性能往往较差，这是由于比较相似的应用容易被相互错分所致，例如同属于社交应用的 QQ 和微信。某些文献**错误!未找到引用源。**采用粗粒度的互联网流量分类，将相似的应用归为一类，但现有的应用划分方式主要基于主观判断。本节将 MFEC 方法应用到 MAD 数据集，分析各应用之间的相似度，为应用划分和用户行为分析提供客观建议。

1）App 相似度评价指标

基于 Jaccard 距离指标，本文提出 App 间相似度指标，其定义如下：

$$similarity(A,B)=\frac{\sum_{i=1}^{n}\frac{min\{a_i,b_i\}}{a_i+b_i}}{n} \tag{7}$$

其中:a_i 和 b_i 分别表示应用 A 和应用 B 在第 *i* 个聚簇中的样本数，聚簇总数为 *n*。此指标的取值范围为[0, 0.5]。当取值为 0，表示 A 和 B 没有样本聚类到相同的簇，两者相似度最低；当取值为 0.5，表示 A 和 B 之间相似度最高。

2）App 相似度的分析

MFEC 聚类后的 App 相似度矩阵如表 4 所示，灰体表示自身的相似度，粗体表示其相似度高于相似度阈值（所有相似度的平均，即 0.278）。其中，某些 App（如 Browser）与多个 App 相似，而某些 App（YahooMail）则没有与之相似的 App。为了更清晰的表明 App 的相似关系，本文建立了如图 8 所示的关联图。在图 8 中，每个节点表示矩阵中行和列的某个 App，若两个 App 相似度大于某个阈值，则建立一条边，并且边的权重为相似度。

表 4　MAD 数据集 App 相似度矩阵

	Browser	JdShop	MgTV	QQ	VipShop	WeChat	Weibo	YahooMail	Youku
Browser	0.500	0.342	0.224	0.249	0.334	0.347	**0.364**	**0.316**	0.254
JdShop	0.342	0.500	0.186	0.306	**0.349**	0.318	0.307	0.209	0.274
MgTV	0.224	0.186	0.500	0.219	0.195	0.235	0.158	0.236	**0.339**
QQ	0.249	0.306	0.219	0.500	0.318	**0.400**	0.247	0.304	0.257
VipShop	0.334	**0.349**	0.195	0.318	0.500	0.307	0.254	0.284	0.182
WeChat	0.347	0.318	0.235	**0.400**	0.307	0.500	0.324	0.290	0.307
Weibo	**0.364**	0.307	0.158	0.247	0.254	0.324	0.500	0.245	0.314
YahooMail	0.316	0.209	0.236	0.304	0.284	0.290	0.245	0.500	0.213
Youku	0.254	0.274	**0.339**	0.257	0.182	0.307	0.314	0.213	0.500

图 8 表明，与 WeChat 相似的 App 有 7 个，按相似度大小排列分别是 QQ、Browser、Weibo、JdShop、Youku、Vipshop 和 YahooMail。类似地，与 Weibo 相似的有 Browser、WeChat、JdShop 和 Youku。其中，形成多个全关联子图，例如：WeChat、Weibo、JdShop 和 Browser 等。全关联子图的 App 可划分为一类。若不断将相似阈值提高，则可以找出与某个 App 最为相似的 App，帮助提高 App 划分的准确率。例如当相似阈值提高到 0.32 时，仅有 WeChat、Browser、Weibo 形成全关联子图，即表示这三个应用最为相似，可建议将它们划分为一个类别。这也意味着，在此实验数据集上，用户在使用 WeChat 和 Weibo 方面，更多的表现为浏览行为。

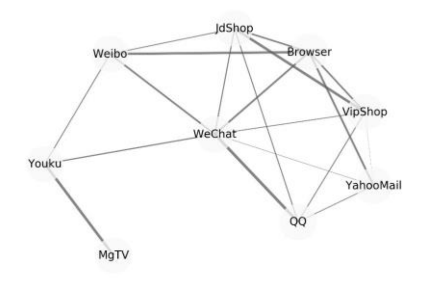


图 8　MAD 数据集上 App 相似关联图

4　结束语

本文基于多项评估指标，分析 K-均值和谱聚类方法在不同特征集合或不同类别标签的移动互联网流量数据集上的聚类性能。实验结果表明，K-均值在 App 流量识别方面的性能优于谱

聚类，并且单向流特征和双向流特征更适用于 App 流量识别。为了综合利用不同角度的特征集合，本文提出基于多个特征集合的集成聚类方法 MFEC，提高聚类性能。实验结果表明，MFEC 能进一步提高 K-均值的 App 识别准确率。最后，本文将 MFEC 方法运用于 App 相似度分析，分析结果可辅助于用户 App 上网行为的分析，并为繁杂的 App 归类提供客观的建议。本文主要对常用的特征集合进行了分析与比较，未来将结合 MFEC 方法，在移动互联网流量数据集上研究性能更优的特征集合。

参考文献：

[1] 田旭. 互联网流量识别技术研究 [D]. 北京: 北京邮电大学, 2012. (Tian Xu. Research on Internet traffic identification technology [D]. Beijing: Beijing University of Posts and Telecommunications, 2012.)

[2] 林森, 徐鹏, 刘琼. 基于支持向量机的 Internet 流量分类研究 [J]. 计算机研究与发展, 2009, 46 (3): 407-414. (Lin Sen, Xu Peng, Liu Qiong. Traffic classification based on support vector machine [J]. Journal of Computer Research and Development, 2009, 46 (3): 407-414.)

[3] Soysal M, Schmidt E G. Machine learning algorithms for accurate flow-based network traffic classification: evaluation and comparison [J]. Performance Evaluation, 2010, 67 (6): 451-467.

[4] Peng Lizhi, Yang Bo, Chen Yuehui. Effective packet number for early stage internet traffic identification [J]. Neurocomputing, 2015, 156 (C): 252-267.

[5] Liu Zhen, Wang Ruoyu, Tao Ming, *et al.* A class-oriented feature selection approach for multi-class imbalanced traffic datasets based on local and global metrics fusion [J]. Neurocomputing, 2015, 168 (2015): 365-381.

[6] Zhang Jun, Chen Xiao, Xiang Yang, *et al.* Robust network traffic classification [J]. IEEE//ACM Trans on Networking, 2014, 23 (4): 1257-1270.

[7] Aceto G, Ciuonzo D, Montieri A, *et al.* Multi-classification approaches for classifying mobile App traffic [J]. Journal of Network & Computer Applications, 2017, 103: 131-145.

[8] Park B, Hong J W, Won Y J. Toward fine-grained traffic classification [J]. IEEE Communications Magazine, 2011, 49 (7): 104-111.

[9] Erman J, Arlitt M, Mahanti A. Traffic classification using clustering algorithms [C]// Proc of ACM SIGCOMM Workshop On Mining Network Data. 2006: 281-286.

[10] Bernaille L, Teixeira R, Salamatian K. Early application identification [C]// Proc of ACM CoNEXT. New York: ACM Press, 2006: 1-6.

[11] 周文刚, 陈雷霆, 董仕. 基于谱聚类的网络流量分类识别算法 [J]. 电子测量与仪器学报, 2013, 27 (12): 1114-1119. (Zhou Wengang, Chen Leiting, Dong Shi. Network traffic classification algorithm based on spectral clustering [J]. Journal of Electronic Measurement and Instrument, 2013, 27 (12): 1114-1119.)

[12] 鲁刚, 余翔湛, 张宏莉, 等. 基于集成聚类的流量分类架构 [J]. 软件学报, 2016, 27 (11): 2870-2883. (Lu Gang, Yu Xiangzhan, Zhang Hongli, *et al.* Traffic classification framework based on integrated clustering [J]. Journal of Software, 2016, 27 (11): 2870-2883.)

[13] 何震凯, 阳爱民, 刘永定, 等. 一种使用DBSCAN聚类的网络流量分类方法 [J]. 计算机应用研究, 2009, 26 (9): 3461-3464. (He Zhenkai, Yang Aimin, Liu Yongding, *et al.* A network traffic classification method using DBSCAN clustering [J]. Application Research of Computers, 2009, 26 (09): 3461-3464.)

[14] Tongaonkar A. A look at the Mobile App Identification Landscape [J]. IEEE Internet Computing, 2016, 20 (4): 9-15.

[15] Fu Yanjie, Xiong Hui, Lu Xinjiang, *et al.* Service usage classification with encrypted internet traffic in mobile messaging Apps [J]. IEEE Trans on Mobile Computing, 2016, 15 (11): 2851-2864.

[16] 张馨予. 基于社团结构的移动互联网流量分析与应用识别 [D]. 北京: 北京邮电大学, 2014. (Zhang Xinyu. Mobile Internet traffic analysis and application identification based on community structure [D]. Beijing: Beijing University of Posts and Telecommunications, 2014.)

[17] Tongaonkar A, Keralapura R, Nucci A. Challenges in network application identification [C]// Proc of the 5th USENIX Conference on Large-Scale Exploits and Emergent Threats. 2012: 1-3.

[18] 刘珍, 王若愚, 蔡先发, 等. 互联网流量分类中流量特征研究 [J]. 计算机应用研究, 2017, 34 (1): 8-14, 41. (Liu Zhen, Wang Ruoyu, Cai Xianfa, *et al.* Survey on traffic features in internet traffic classification [J]. Application Research of Computers, 2017, 34 (01): 8-14, 41.)

[19] Moore A, Zuev D, Crogan M L. Discriminators for use in flow-based classification [R]. Queen Mary and Westfield College, 2005: 1-16.

[20] Qin Tao, Wang Lei, Liu Zhaoli, *et al.* Robust application identification methods for P2P and VOIP traffic classification in backbone networks [J]. Knowledge-Based Systems, 2015, 82 (C): 152-162.

[21] Dainotti A, Pescape A, Kim H C. Traffic classification through joint distributions of packet-level statistics [C]// Proc of IEEE Global Telecommunications. 2011: 1-6.

[22] Mongkolluksamee S, Visoottiviseth V, Fukuda K. Enhancing the performance of mobile traffic identification with communication patterns [C]// Proc of Computer Software and Applications Conference. 2015: 336-345.

[23] Liu Zhen, Wang Ruoyu. Mobilegt: A system to collect mobile traffic trace and build the ground truth [C]// Proc of Telecommunication Networks and Applications Conference. 2017: 142-144.

[24] 王变琴, 余顺争. 未知网络应用流量的自动提取方法 [J]. 通信学报, 2014, 35 (07): 164-171. (Wang Bianqing, Yu Shunzheng. Unknown Network Application Traffic Automatic Extraction Method [J]. Journal on Communications, 2014, 35 (7): 164-171.)

[25] Williams N, Zander S, Armitage G. A preliminary performance comparison of five machine learning algorithms for practical IP traffic flow classification [J]. ACM SIGCOMM Computer Communication Review, 2006, 36 (5): 5-16.